Praxis | Spracherkennung lokal c't 14/2023 S. 140

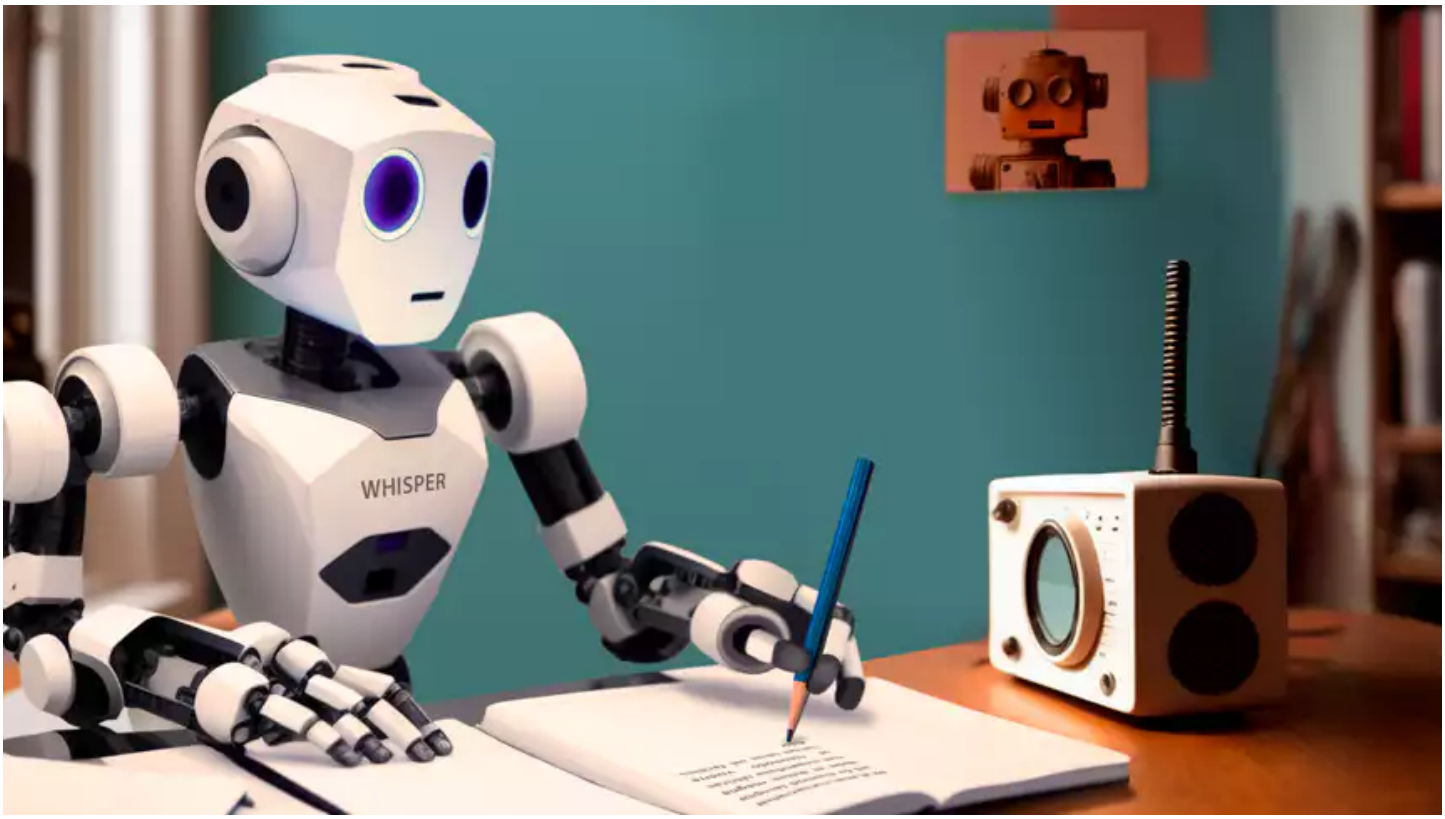

Bild: KI Stable Diffusion | Bearbeitung: c't

Spracherkennung im Eigenbau

Whisper wandelt Gesprochenes in Text um

Die Open-Source-Spracherkennung Whisper transkribiert Sprache aus Audiodateien mit sehr guter Erkennungsquote und versteht sich sogar auf Zeichensetzung. Sie läuft lokal, ein Mittelklasse-PC ist dafür schnell genug. Wir zeigen, wie Sie Whisper einrichten und bedienen.

In Sprache in Videos von Hand zu transkribieren, ist daher ein Knochenjob. Diverse Dienstleister bieten Transkriptionen an, allerdings haben zuverlässige Angebote ihren Preis. Hinzu kommt: Wer seine Audiodaten an einen Anbieter von Spracherkennung schickt, gibt damit eventuell vertrauliche Daten aus der Hand.

Seit September 2022 gibt es Whisper, eine kostenlos nutzbare Transkriptionssoftware des US-amerikanischen KI-Start-ups OpenAI, das mit dem KI-Chatbot ChatGPT Furore gemacht hat. Das Open-Source-Programm analysiert Audioaufzeichnungen und wandelt darin enthaltene Sprache in Textdateien um. Für die Einrichtung nutzen Sie die Kommandozeile.

Whisper beherrscht laut OpenAI 96 Sprachen, Deutsch ist demnach unter den fünf mit der geringsten Fehlerrate bei der Erkennung. Die Sprach-KI arbeitet sich mühelos durch minuten- bis stundenlange Aufzeichnungen, mithilfe der freien Multimediasoftware ffmpeg kann sie nahezu jedes Ton- oder Videoformat verarbeiten. Noch kann Whisper bei Aufnahmen mit mehreren Sprechern nicht zwischen den einzelnen Personen unterscheiden. Doch auch an dieser Aufgabe wird bereits getüftelt – mehr dazu im Ausblick am Ende des Artikels.

Sieht aus wie Sprache

Whisper unterteilt das Audiomaterial in 30-sekündige Abschnitte. Diese wandelt es dann in sogenannte Mel-Spektrogramme um; OpenAI spricht in der Beschreibung von „log-Mel spectrograms“. Das ist eine grafische Darstellungsform, die die Lautstärke der für den Menschen hörbaren Frequenzen im Zeitverlauf darstellt. Die Whisper-KI hat die Worterkennung anhand solcher „Bilder“ trainiert. Das Ergebnis des Trainings ist ein Sprachmodell für die Transkription.

OpenAI bietet fünf unterschiedlich große Modelle an; sie heißen tiny, base, small, medium, large. Das größte Modell bekam kürzlich ein Update. large-v2 ist ebenso wie der Vorgänger large etwa 3 GByte groß. Das kleinste Modell base belegt lediglich 139 MByte Festspeicher. Whisper lädt das Sprachmodell beim ersten Verwenden automatisch herunter.

Das Programmpaket kommt ohne grafische Bedienoberfläche. Es nutzt die Skriptsprache Python und das KI-Framework PyTorch. Wer Python beherrscht, kann sich eigene Ausgabeformate stricken. Für die meisten Anwendungszwecke genügt aber der schnelle Aufruf per Kommandozeile. Standardmäßig gibt Whisper Klartext, drei Untertitelformate mit Zeitmarken, ein Tabellenformat (TSV) und eine JSON-Protokolldatei aus.

KI-typisch hängt die Dauer der Verarbeitung von der Grafikkarte des Rechners ab [1]. Schnelle Ergebnisse liefern PCs mit leistungsfähiger Nvidia-Grafikkarte. Verwendet man für die höchste Erkennungsqualität das größte Sprachmodell, setzt Whisper etwa 10 GByte VRAM voraus.

Vorbereitung

Whisper läuft unter Windows, macOS und Linux. In den drei Textkästen erklären wir die Besonderheiten unter jedem der drei Betriebssysteme. Auf Windows sowie macOS müssen Sie zunächst Python installieren. Am einfachsten gelingt das mit dem jeweiligen Paketmanager, etwa Chocolatey (Windows) oder Homebrew (macOS). Der erleichtert die Einrichtung notwendiger Komponenten, beispielsweise der Versionsverwaltung git. Wir haben Whisper mit Python in Version 3.9 ausprobiert; mit dieser Version wurde die Transkriptionssoftware trainiert. Laut OpenAI funktioniert sie auch unter 3.8 bis 3.11.

Es empfiehlt sich, für Python und Whisper eine virtuelle Umgebung einzurichten. Auf diese Weise kapseln Sie die Whisper-KI, die dazugehörige Python-Installation und zusätzliche Module (unter anderem pytorch und tiktoken) in einem Ordner. So schließen Sie Inkompatibilitäten mit anderen Python-Projekten aus. Wie Sie Python installieren und für virtuelle Umgebungen vorbereiten, haben wir in c't 5/2022 beschrieben [2].

Einrichten unter Windows

Windows benötigt zunächst eine Paketverwaltung, die git, ffmpeg und python installiert. Dafür empfiehlt sich die Community-Version von Chocolatey, die Privatanwender kostenlos nutzen dürfen. Öffnen Sie die WindowsPowershell und geben folgende Befehle ein, jede der insgesamt drei Zeilen von der Eingabetaste gefolgt:

```
Get-ExecutionPolicy
Set-ExecutionPolicy AllSigned
Set-ExecutionPolicy Bypass -Scope Process -Force; [System.Net.ServicePointMana
```

Wenn nach kurzer Zeit die Eingabeaufforderung erneut erscheint, geben Sie `choco -?` ein. Erscheinen Tipps zur Bedienung von Chocolatey, hat es geklappt. Nun können Sie die relevanten Pakete mit `choco install ffmpeg git python39` einrichten. Das braucht einen Moment.

Wartung und löschen

Chocolatey überprüft nicht ohne Ihre Initiative, ob es neue Versionen gibt. Um alle installierten Pakete auf einen Rutsch zu aktualisieren, geben Sie `choco upgrade all -y` ein. Falls Sie Whisper nicht mehr benötigen und alle zusätzlichen Installationen loswerden wollen, löscht der Befehl `choco uninstall all` sämtliche Pakete, in diesem Fall ffmpeg, git und py-

Einrichten unter Linux

Linux-Distributionen sind bestens für Whisper geeignet, denn sie bringen bereits eine aktuelle Python-Version sowie eine Paketverwaltung mit. Ubuntu 22.04 etwa verwendet `apt` zum Installieren von Paketen und Modulen; Python liegt bereits in Version 3.10 vor. Starten Sie zunächst das Kommandozeilenprogramm Ihrer Distribution und geben Sie `sudo apt install ffmpeg git` ein. Nach Bestätigen mit Enter fragt das System nach Ihrem Administratorpasswort. Nachdem Sie es eingegeben haben, werden die Softwarepakete ffmpeg und git eingerichtet.

Ältere Pythonversionen erscheinen nicht in den Paketquellen, weshalb Sie ein zusätzliches Softwareverzeichnis hinzufügen. Falls Sie Ubuntu 22.04 oder ein darauf basierendes Derivat verwenden, geben Sie im Terminal folgende Zeilen ein, die Sie jeweils mit Enter bestätigen:

```
sudo add-apt-repository ppa:deadsnakes/ppa
sudo apt install python3.10-venv
sudo apt install python3.9
sudo apt install python3.9-venv
```

Wartung und löschen

Falls Ihre Distribution nicht von selbst auf Updates prüft und sie zur Installation anbietet, sollten Sie regelmäßig `sudo apt update` gefolgt von `sudo apt upgrade` verwenden, jeweils bestätigt mit der Enter-Taste und gegebenenfalls Ihrem Administratorkennwort. Falls Sie Ihre Whisper-Installation wieder entfernen möchten, löscht `sudo apt remove python3.9 ffmpeg git` die installierten Pakete. `sudo add-apt-repository -r ppa:deadsnakes/ppa` entfernt das Python-Depot aus den Paketquellen.

Einrichten unter macOS

Um Whisper auf dem Mac zu nutzen, benötigen Sie zunächst die Kommandozeilenwerkzeuge der Entwicklungsumgebung Xcode. Wenn Sie diese über den Mac App Store installieren, belegt sie etwa 40 GByte Festspeicher. Wer nicht plant, eigene Apps zu entwickeln, kann Platz sparen: Mit dem Kommandozeilenbefehl `xcode-select --install` lösen Sie lediglich die Installation der essenziellen Softwarepakete aus; deren Platzbedarf liegt unter 2

```
/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/
```

Anschließend folgen Sie den Anweisungen im Kommandozeilenfenster. Das Ganze dauert vielleicht eine Minute. Danach geben Sie folgende Befehle ein, um ffmpeg und Python 3.9 zu installieren:

```
brew update
brew install ffmpeg python@3.9
```

Wartung und löschen

Die per Homebrew installierten Softwarepakete halten Sie auf dem jeweils aktuellen Stand, indem Sie regelmäßig `brew upgrade` in der Kommandozeile eingeben. Wenn Sie Ihr Whisper-Experiment beenden wollen, deinstallieren Sie die beiden Pakete über `brew uninstall ffmpeg python@3.9` Die Paketverwaltung selbst entfernen Sie mit dem Befehl `/bin/bash -c "$(curl -fsSL https://raw.githubusercontent.com/Homebrew/install/HEAD/uninstall.sh)"`

Virtuelle Umgebung einrichten

Nachdem Sie die für Ihr Betriebssystem spezifische Vorbereitung abgeschlossen haben, geht es für alle Systeme in gleicher Form weiter, auf der Kommandozeile. Suchen Sie sich einen Ort auf Ihrer Dateiverwaltung, den Sie als Spielwiese für Ihre Whisper-Experimente nutzen wollen. Verwenden Sie dafür die Dateiverwaltung Ihres Betriebssystems. Wir erschufen einen Ordner namens „Skripte“ im Benutzerordner; die folgenden Beschreibungen gehen von diesem Ordner als Ausgangspunkt aus. Auf der Kommandozeile wechseln Sie mit dem Befehl `cd Skripte` dorthin.

Hier machen Sie es der KI gemütlich, indem Sie ihr eine virtuelle Umgebung mit Python 3.9 einrichten. Der Befehl `python3.9 -m venv whisperenv` erzeugt einen Ordner namens „whisperenv“, der Befehl `source whisperenv/bin/activate` setzt die virtuelle Umgebung in Gang. Unter Windows lautet der Befehl `Skripte\whisperenv\Scripts\activate.bat`

Am Anfang der Eingabeaufforderung erscheint nun der Name Ihrer virtuellen Umgebung. Falls Sie Python in anderen Projekten verwenden und darum whisperenv verlassen wollen, gelingt das mit dem Befehl `deactivate`. Die virtuelle Umgebung ist nur im aktuellen Kommandozeilenfenster aktiv. Wenn Sie also den Rechner neu starten, das Kommandozeilenprogramm zwischenzeitlich beendet haben oder in einem neuen, zweiten Terminal-Fenster mit Whisper arbei-

über:

```
pip3 install git+https://github.com/openai/whisper.git
```

Dabei sucht sich `pip` alle Pakete, die es benötigt, und installiert sie gleich mit. Und hier macht sich die virtuelle Umgebung bezahlt: Alles, was Sie fortan mit `pip` installieren, landet im Ordner ~/Skripte/whisperenv/bin.

Überprüfen Sie mittels `whisper --help`, ob die Kommandozeile das installierte Paket findet. Unter macOS meldete sie bei unseren Tests häufiger, sie kenne kein Whisper. In diesem Fall helfen Sie ihr mit:

```
sudo ln -s ~/Skripte/whisperenv/bin/whisper /usr/bin
```

auf die Sprünge; die Tilde ~ ist eine Abkürzung für das Home-Verzeichnis des gerade angemeldeten Nutzers.

Sämtliche genannten Kommandozeilenbefehle finden Sie – nach Betriebssystemen sortiert – unter ct.de/ypvb als Textdatei zum Herunterladen.

Die erste Transkription

Jetzt ist alles startklar für eine erste Transkription. Zu Beginn belegt der Orderinhalt nur etwa ein halbes GByte Speicherplatz, doch das ändert sich schlagartig beim ersten Durchlauf, in dem die Parameter des Sprachmodells heruntergeladen werden.

Standardmäßig legt das Kommandozeilenprogramm die bis zu 3 GByte große Modell-Datei unter `~/.cache/whisper/` ab, wo sie beim Aufräumen schwer aufzuspüren ist. Damit alles schön sortiert bleibt, legen Sie sich im Ordner whisperenv Ihrer virtuellen Umgebung einen Unterordner namens models an. Mit der Option `--model_dir` geben Sie Whisper mit auf den Weg, wo es das Modell ablegen soll und später wiederfindet.

Verwenden Sie für den ersten Testlauf am besten eine ein- oder zweiminütige Aufzeichnung, um den Zeitfaktor festzustellen, den Ihr Rechner für die Audiotranskription benötigt. Speichern Sie sie in dasselbe Verzeichnis, in dem der whisperenv-Ordner liegt, in unserem Fall der Skripte-Ordner. Dann geben Sie den folgenden Befehl ein, wobei Sie „Audiodatei.wav“ durch den Dateinamen Ihrer Aufzeichnung inklusive Dateiendung ersetzen:

```
whisper --model_dir whisperenv/models Audiodatei.wav
```

Sobald Sie dies mit der Eingabetaste bestätigen, lädt Whisper zunächst das Sprachmodell her-

deutscher Sprache allerdings nicht sonderlich zuverlässig transkribiert. Eine Nvidia-GTX 1060-Karte verschriftlichte die gesprochene Minute in rapiden 17 Sekunden.

Die Kommandozeile zeigt interaktiv den Fortschritt der Transkription.

Im Kommandozeilenfenster können Sie Whisper beim Transkribieren zusehen. Die erkannten Textzeilen erscheinen dort mit Zeitmarken. Wenn die Transkription vollständig ist, gibt es fertige Textdateien: In der Standardeinstellung schreibt Whisper fünf Dateien ins aktuelle Verzeichnis. Wenn Sie unserem Beispiel gefolgt sind, entdecken Sie sie im Ordner „Skripte“.

Die TXT-Datei enthält die reine erkannte Sprache ohne zusätzliche Informationen. Die Dateien mit Endung SRT, VTT sowie TSV enthalten zudem Zeitmarken in der jeweiligen Nomenklatur. SRT und VTT sind Untertitelformate, die Sie auf YouTube und ähnlichen Portalen als Meta-Information zu Ihrem Video hochladen können. TSV (Tab-Separated Values) ist ein Klartext-Tabellenformat. Die JSON-Datei enthält zusätzlich Informationen zum Dekodierungsverlauf. Wenn Sie nur das Untertitelformat SRT haben wollen, ergänzen Sie den Kommandozeilenbefehl mit der entsprechenden Option: `whisper […] --output_format srt`. Mit `--output_dir` geben Sie einen Zielordner an; standardmäßig landen die Textdateien im aktuellen Arbeitsverzeichnis der Kommandozeile, bei uns also im Ordner „Skripte“.

dell, die Erkennungsqualität steigt jedoch sprunghaft. Falls Sie die Rechenpower oder Zeit haben, nutzen Sie am besten large-v2. Dann müssen Sie am Text nur wenig verbessern und bei den Satzzeichen lediglich Gedanken-, Bindestriche und Anführungszeichen ergänzen. Wo Sätze enden und Kommata hingehören, erkennt Whisper erstaunlich zielsicher.

Für das large-v2-Modell benötigen Sie allerdings eine Nvidia-Grafikkarte mit mindestens 10 GByte VRAM. Wenn Whisper über zu wenig Speicher meckert, aber Ihr Rechner mehr als 10 GByte RAM hat, weisen Sie mit der Option `--device 'cpu'` die Aufgabe dem Prozessor zu; das dauert allerdings gut zehnmal so lang. Wer einen neueren Mac mit M1- oder M2-Prozessor nutzt, kann die Aufgabe mit dem Parameter `mps` eventuell beschleunigen.

Über zusätzliche Parameter erleichtern Sie Whisper die Arbeit: `--language 'de'` weist die KI an, sich vom Start weg auf deutsche Sprache festzulegen. Tut sich Whisper an einer Aufnahme schwer, helfen Sie der KI mit einer manuellen Transkription der ersten 30 Sekunden auf die Sprünge. Über den Parameter `--initial_prompt '[Transkript]'` geben Sie Whisper diese Starthilfe mit auf den Weg.

Die Option `--task translate` übersetzt eine Aufzeichnung direkt ins Englische. Die Qualität ist eher mittelmäßig, spezialisierte KIs wie DeepL translate erzeugen weitaus überzeugendere Übersetzungen. Die Whisper-Übersetzung reicht aber aus, um auf die Schnelle die Kernaussagen einer fremdsprachigen Audiodatei zu entschlüsseln.

Whisper-Optionen

Option	Parameter	Beschreibung
--model	tiny, base, **small**, medium, large, large-v2	entscheidet über die Erkennungsgenauigkeit
--model_dir	[Pfad] ~/cache./whisper	legt Speicherort für die Modell-Dateien fest
--output_format	txt, vtt, tsv, srt, json, **all**	legt das Ausgabeformat fest
--output_dir	[Pfad]	legt den Zielort für die Ausgabe fest
--task	**transcribe**, translate	transkribiert oder übersetzt ins Englische
--initial_prompt	[Text]	erzeugt ein Transkript der ersten 30 Sekunden
--device	**cuda**, cpu, mps	Chipzuweisung

Standardparameter sind **fett** hervorgehoben

Speichern als Skript

dieselbe Weise. Öffnen Sie einen Texteditor und geben Sie folgenden Inhalt ein:

```
#!/bin/bash
whisper --model_dir '~/Skripte/whisperenv/models' --language 'de'--output_dir '~
```

Diese Datei sichern Sie mit dem Namen „transkribiere“ (ohne Dateiendung) im Unterordner „bin“ des whisperenv-Ordners. In der Kommandozeile sorgen Sie mit dem folgenden Befehl dafür, dass dieses Skript ausführbar wird:

```
chmod +x ~/Skripte/whisperenv/bin/transkribiere
```

Fortan geben Sie in der Kommandozeile `transkribiere` (mit Leerzeichen am Ende) ein und ziehen anschließend eine Video- oder Audiodatei per Drag & Drop auf das Terminalfenster. Das Programm erzeugt dann automatisch den passenden Textpfad zu dieser Datei. Sie müssen nur noch die Enter-Taste betätigen, um die Transkription zu starten.

Mac-Anwender können sich mit der Kurzbefehle-App eine Abkürzung anlegen, in der die Aktion „Shell-Skript ausführen“ das Skript aufnimmt. Im Skript müssen Sie die virtuelle Umgebung starten. Wenn Sie „Im Share-Sheet anzeigen“ aktivieren, erscheint der Kurzbefehl dann systemweit im Teilen-Menü.

Die Kurzbefehle-App erleichtert Mac-Nutzern die Whisper-Transkription aus der grafischen

```
pip3 install --upgrade --no-deps --force-reinstall git+https://github.com/openai
```

Dies installiert das Whisper-Modul neu. Um die zusätzlich installierten Module aktuell zu halten, fügen Sie per `pip install pip-review` ein Python-Tool hinzu. Danach überprüfen Sie mit dem Befehl `pip-review --interactive`, ob für Ihre Python-Softwaresammlung neue Versionen bereitstehen. Für jedes Paket entscheiden Sie einzeln, ob Sie dies installieren wollen.

Der whisperenv-Ordner wächst mit der Zeit wahrscheinlich auf einige GByte Größe an, je nachdem, welche Modelle Sie herunterladen. Falls Sie sich nach einigen Experimenten doch gegen Whisper entscheiden, löschen Sie einfach den gesamten Ordner. Um auch Paketverwaltung, python, ffmpeg und gegebenenfalls git zu entfernen, folgen Sie den Anweisungen im Kasten Ihres Betriebssystems.

Was versteht Whisper?

In umfangreichen Experimenten kristallisierten sich bei uns einige Erfahrungswerte heraus. Beim voreingestellten Modell small weisen deutschsprachige Transkriptionen noch recht viele Fehler auf, wir raten davon ab. Bei medium geht die Nachkorrektur leicht von der Hand, mit large_v2 gibt es kaum noch etwas zu verbessern.

Während beim base-Modell noch etliche Fehler (farbig markiert) in der Transkription erscheinen, …

Die Erkennungsrate hängt zudem von der Aufnahmequalität ab: Je mehr Störgeräusche, desto mehr Fehler schleichen sich ein. Andererseits kommt Whisper mit gänzlicher Stille auch nicht gut zurecht. Die KI halluziniert dann Inhalte oder wiederholt den letzten erkannten Satz mehrfach.

... erzeugt Whisper mit dem medium-Modell Text mit deutlich weniger Fehlern (Quelltext links, Transkript rechts).

Ausblick

Der Entwickler OpenAI veröffentlichte Whisper als Open-Source-Projekt unter der MIT-Lizenz. Das erlaubt es Entwicklern, die Sprachmodelle und Algorithmen recht frei in eigene Projekte zu integrieren. Faster-Whisper will schneller und ressourcenschonender transkribieren; das Python-Projekt verwendet CTranslate2 anstelle von Torch für die Berechnung. whisper.cpp setzt die Spracherkennung in der Programmiersprache C++ um. Diese Variante strebt an, mehr Grafikchips zu unterstützen und den Bedarf an (V)RAM zu reduzieren.

Wer Aufnahmen mit mehreren Stimmen – etwa Interviews oder Podcasts – verschriftlichen will, wünscht sich eine automatische Stimmzuordnung (Diarization). Das GitHub-Projekt pyannote-whisper arbeitet daran, die Whisper-Texterkennung mit der Stimmenzuordnungs-KI PyAnnote Audio zusammenzubringen. Dies ist aktuell allerdings noch weit entfernt von einer zuverlässigen Verwendung; beim Installieren gibt es zudem etliche Hürden zu meistern.

Die Dynamik im KI-Umfeld ist hoch. Vielleicht zeichnen sich bereits in wenigen Wochen auch für dieses und weitere Einsatzszenarios sinnvolle Lösungen ab. Transkripte und Untertitelspuren lassen sich mit Whisper schon jetzt weitaus schneller anfertigen, als es beim Schreiben von Hand möglich wäre – mit der hier vorgestellten Methode auch lokal auf dem eigenen Rechner. *(dwi@ct.de)*

Literatur

[1] Carsten Spille, Grips-Chips, Die besten Grafikkarten für Stable Diffusion AI, c't 9/23, S. 64

[2] Ronald Eikenberg, Jahn Mahn: Draufgebeamt, So richten Sie Python schnell und einfach ein, c't 5/2022, S. 20

Downloads, Kommandozeilenbefehle: *ct.de/ypvb*

```
sudo add-apt-repository ppa:deadsnakes/ppa.
```

Kommentare lesen (8 Beiträge)